Bayesian model evidence as a practical alternative to deviance information criterion

ROYAL SOCIETY OPEN SCIENCE

rsos.royalsocietypublishing.org

Research

Cite this article: Pooley CM, Marion G. 2018 Bayesian model evidence as a practical alternative to deviance information criterion. *R. Soc. open sci.* **5**: 171519.
http://dx.doi.org/10.1098/rsos.171519

Received: 5 October 2017
Accepted: 13 February 2018

Subject Category:
Mathematics

Subject Areas:
mathematical modelling/applied mathematics/computational biology

Keywords:
Bayes' factor, Bayesian model evidence, marginal likelihood, Markov chain Monte Carlo, thermodynamic integration, deviance information criterion

Author for correspondence:
C. M. Pooley
e-mail: christopher.pooley@roslin.ed.ac.uk

Electronic supplementary material is available online at https://dx.doi.org/10.6084/m9.figshare.c.4020832.

THE ROYAL SOCIETY PUBLISHING

Bayesian model evidence as a practical alternative to deviance information criterion

C. M. Pooley[1,2] and G. Marion[2]

[1]The Roslin Institute, The University of Edinburgh, Midlothian EH25 9RG, UK
[2]Biomathematics and Statistics Scotland, James Clerk Maxwell Building, The King's Buildings, Peter Guthrie Tait Road, Edinburgh EH9 3FD, UK

CMP, 0000-0002-8779-4477

While model evidence is considered by Bayesian statisticians as a gold standard for model selection (the ratio in model evidence between two models giving the Bayes factor), its calculation is often viewed as too computationally demanding for many applications. By contrast, the widely used deviance information criterion (DIC), a different measure that balances model accuracy against complexity, is commonly considered a much faster alternative. However, recent advances in computational tools for efficient multi-temperature Markov chain Monte Carlo algorithms, such as steppingstone sampling (SS) and thermodynamic integration schemes, enable efficient calculation of the Bayesian model evidence. This paper compares both the capability (i.e. ability to select the true model) and speed (i.e. CPU time to achieve a given accuracy) of DIC with model evidence calculated using SS. Three important model classes are considered: linear regression models, mixed models and compartmental models widely used in epidemiology. While DIC was found to correctly identify the true model when applied to linear regression models, it led to incorrect model choice in the other two cases. On the other hand, model evidence led to correct model choice in all cases considered. Importantly, and perhaps surprisingly, DIC and model evidence were found to run at similar computational speeds, a result reinforced by analytically derived expressions.

1. Introduction

The advent of Markov chain Monte Carlo (MCMC) has enabled Bayesian inference for increasingly complex models [1], but much of this progress has been in the context of single model inference. Model selection refers to the problem of selecting or weighting different models in the light of the available data. One of the key challenges is to avoid overfitting the

data by selecting unduly complex models. More reliable and efficient model selection is critical to the wider adoption of model-based scientific discovery, especially for applications in dynamic process-based modelling. Multi-model Bayesian inference automatically, and implicitly, includes a penalty for unnecessary model complexity, and thus guards against overfitting [2,3]. Reversible jump MCMC, introduced by Green [4], in principle can be used to implement Bayesian model choice (this estimates the model posterior probability by the proportion of samples the MCMC chain spends within that model). However, this approach can be difficult to implement in practice [5].

From a Bayesian perspective, deviance information criterion (DIC) is an approximate model selection method which tries to explicitly balance model complexity with fit to the data [6]. However, there are increasing concerns with regard to its discriminatory performance [7], particularly in the presence of latent variables where there is no unique definition [8]. The practical issues that arise in the implementation of fully Bayesian approaches to model choice are illustrated by the fact that, for example, DIC is the only model selection tool in widespread use for assessing the fit of dynamic stochastic epidemiological models [9–11]. The aim of this paper is to point out the potential unreliability of DIC and to show that Bayesian model selection can be undertaken in a statistically consistent way that is not hard to implement and not substantially computationally slower (and is actually faster in some cases).

We focus on stochastic models that contain parameters θ and latent variables x. To take an epidemiological example, the parameters describe the dynamics of the system (e.g. the average infection and recovery rates of individuals) and the latent variables are the unobserved consequences of those dynamics (e.g. the infection and recovery events). The model latent space behaviour is characterized by a 'latent process likelihood' $\pi(x|\theta)$. We assume that an 'observed data likelihood' $P(y|\theta,x)$ describes the probability of some observed data y, given a particular set of latent variables x and model parameters θ.[1] If we define a prior distribution $\pi(\theta)$, then Bayes' theorem enables the posterior distribution to be expressed as

$$P(\theta, x \mid y) = \frac{P(y \mid \theta, x)\pi(x \mid \theta)\pi(\theta)}{P(y)}, \tag{1.1}$$

where the normalizing factor

$$P(y) = \int P(y \mid \theta, x)\pi(x \mid \theta)\pi(\theta)\, \mathrm{d}\theta\, \mathrm{d}x \tag{1.2}$$

is known as the model evidence, or marginal likelihood [12]. Note, in the above equation, the dependence on the particular model m is implicit. When the prior belief ascribed to each model is equal, this measure forms the sole basis of Bayesian model selection, i.e. the ranking of competing models in the light of the data, with the highest evidence identifying the best model. The famous Bayes' factor [13] comparing models m_1 and m_2 is simply the ratio of the evidence given by data y to model m_1 relative to that given by model m_2:

$$B_{1,2} = \frac{P(y \mid m_1)}{P(y \mid m_2)}. \tag{1.3}$$

A Bayes factor of 10 is typically considered to represent strong evidence favouring m_1 over m_2 [14].

Recently, a variety of techniques have been developed for calculating the model evidence either exactly, e.g. through annealed importance sampling (AIS) [15] or steppingstone sampling (SS) [16], or approximately, e.g. using thermodynamic integration [17–20]. These approaches enable evidence calculation one model at a time and thus represent a practical alternative to reversible jump MCMC for Bayesian model selection.[2] This paper uses a version of SS, as described in the next section. Section 3 introduces DIC and its relationship to model evidence and §4 presents analytical estimates for the computational efficiency of the two approaches. In §5, both the accuracy and efficiency of DIC and SS are assessed using three benchmark models: linear regression models, mixed models (which contain latent random effects) and Markovian stochastic compartmental models widely used to describe epidemic processes. Finally, conclusions are drawn in §6.

2. Model evidence using steppingstone sampling

SS, introduced by Xie *et al*. [16], calculates the model evidence by means of generating samples from K separate MCMC chains at different inverse temperatures $\phi_1 = 1 > \phi_2 > \cdots > \phi_K = 0$. Each chain

[1]Typically, the components of the parameter vector θ that control the model dynamics are distinct from those that parametrize the observed data likelihood.

[2]At least 'practical' when the number of models to choose among is relatively small, or when techniques such as sequential model selection are employed, which can miss the 'best' model.

undergoes changes as a result of proposals that are either accepted or rejected. These proposals can take a variety of forms (e.g. random walk [1], Gibbs sampling [21], Metropolis-adjusted Langevin algorithm [22] etc.) and must be selected such that the MCMC chain can, in principle at least, explore the entirety of parameter and latent variable space. A key difference between SS and ordinary MCMC is that the proposals for each chain k use a Metropolis–Hastings acceptance probability modified by the chain's inverse temperature ϕ_k:

$$\left\{\left(\frac{P(y\mid\theta_p,x_p)}{P(y\mid\theta_i^k,x_i^k)}\right)^{\phi_k}\frac{\pi(x_p\mid\theta_p)\pi(\theta_p)\,j_{p\to i}}{\pi(x_i^k\mid\theta_i^k)\pi(\theta_i^k)\,j_{i\to p}},1\right\}, \tag{2.1}$$

where θ_i^k, x_i^k represents the current sample (indexed by i) and θ_p, x_p represents a proposal (with probability $j_{i\to p}$). If the proposal is accepted then the next sample on the chain is set to $\theta_{i+1}^k=\theta_p, x_{i+1}^k=x_p$, otherwise $\theta_{i+1}^k=\theta_i^k, x_{i+1}^k=x_i^k$. Repeated application of equation (2.1) on each of the chains generates samples distributed in proportion to

$$P(y\mid\theta,x)^{\phi_k}\pi(x\mid\theta)\pi(\theta). \tag{2.2}$$

Consequently, the chain $\phi_1=1$ samples from the posterior (see equation (1.1)), $\phi_K=0$ samples from the prior[3] and chains in between these two extremes provide steppingstones going from one to the other. An unbiased approximation to the model evidence is then given by [16]

$$\hat{P}(y)=\prod_{k=2}^{K}\left(\frac{1}{N}\sum_{i=1}^{N}P(y\mid\theta_i^k,x_i^k)^{\phi_{k-1}-\phi_k}\right), \tag{2.3}$$

which converges on the true model evidence as the number of samples N tends to infinity (see electronic supplementary material, appendix A, for a derivation of this expression).

SS can be illustrated by means of figure 1, which shows $K=6$ chains. The solid line in figure 1*a* shows the variation in the mean of the log of the observed data likelihood as a function of inverse temperature. Note, it decreases from right (posterior) to left (prior), as would be expected. The vertical dashed lines in this diagram represent the inverse temperatures of the different chains which are chosen in accordance with

$$\phi_k=\left(\frac{K-k}{K-1}\right)^5. \tag{2.4}$$

This commonly used power-law scaling is suboptimal (a better but more complicated adaptive scheme is implemented in Friel *et al*. [19]), but empirically is found to work for a large number of problems. By design, it focuses most chains at low inverse temperatures where changes in the observed data likelihood (and variance) are greatest. Figure 1*b* shows the distributions from which samples are generated. As will be discussed later, algorithm efficiency is relatively insensitive to K, provided K is sufficiently large to allow for a substantial overlap between distributions from adjacent chains. In this study, K is chosen to be 50 to ensure this condition is satisfied for all cases investigated. Note even though 50 chains are being updated instead of just one for DIC, it does not imply that the SS method is 50 times slower. This is because the estimate for the model evidence in equation (2.3) converges more rapidly than the corresponding DIC measures, and so, fewer MCMC iterations are required to obtain a given accuracy (as shown in §4).

This paper runs an implementation of SS in which the K chains are updated in parallel with swapping of states between adjacent chains to improve mixing (for details, see electronic supplementary material, appendix A). It should be noted, however, that in some applications, finite memory restrictions make running a large number of chains problematic. In such cases, a scanning approach may be adopted where samples are generated in a sequential manner.[4]

3. Deviance information criterion

Within classical model fitting, the problem of model selection is achieved by considering two competing notions: firstly, a measure of model fit that promotes selecting more accurate models and, secondly, a measure of model complexity, often represented by the number of parameters p_m in the model. This

[3]To calculate model evidence, the prior must be proper ensuring that chain K remains bounded.

[4]Here, burn-in and sampling are alternated as the inverse temperature jumps down in K steps starting at the posterior ϕ_1 and ending up at the prior ϕ_K.

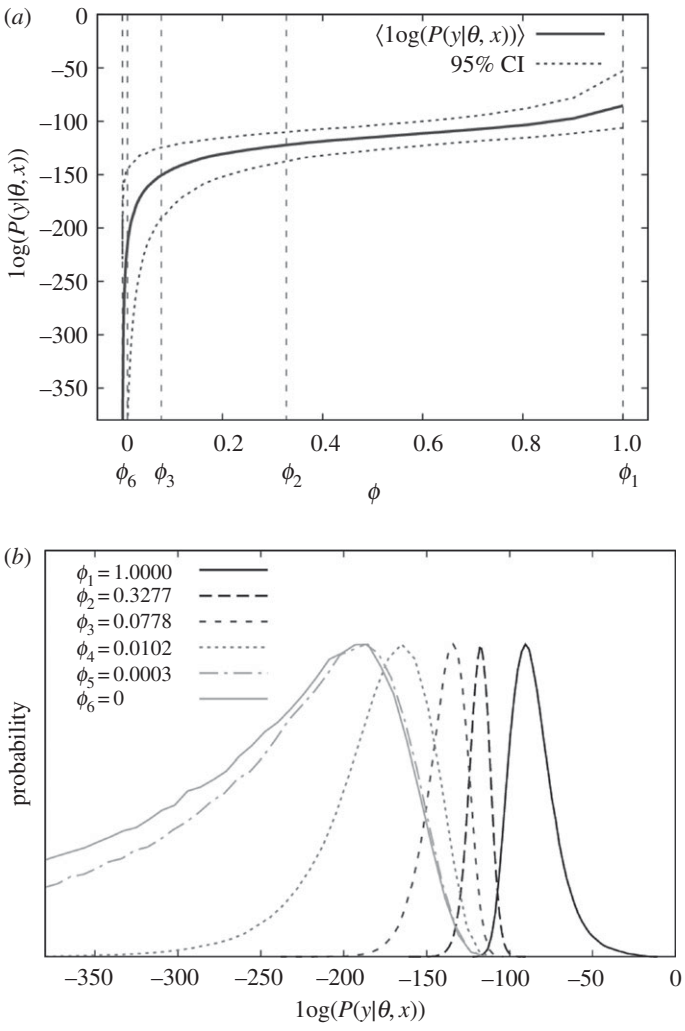

Figure 1. (*a*) A typical example of how the posterior distribution in the log of the observed data likelihood varies as a function of inverse temperature ϕ. This distribution is represented by a mean (solid line) and 95% confidence intervals (denoted by the dotted lines). (*b*) The distributions for $K = 6$ chains from which samples are drawn during SS (with inverse temperatures defined by equation (2.4)).

penalty discourages overfitting (increasing p_m almost always improves goodness of fit). A commonly used measure that balances these two contributions is Akaike's information criterion (AIC) [23]:

$$\mathrm{AIC} = -2\log(P(y \mid \theta_{\max})) + 2p_m, \tag{3.1}$$

where $\theta_{\max}$ is the parameter set that maximizes the observed data likelihood $P(y|\theta)$. Given a number of candidate models, the one with the smallest AIC value is considered best.

For complex hierarchical models, however, establishing p_m is problematic due to a lack of independence between parameters. This has led to the introduction of DIC which uses MCMC results directly. In contrast with SS, however, DIC uses samples from a single posterior chain $k = 1$. As shown below, there are actually multiple definitions for DIC in the literature.

3.1. Deviance information criterion for problems without latent variables

Considering models that do not contain latent variables x, two contrasting definitions for DIC have been proposed. Firstly, Spiegelhalter *et al.* [6] suggested

$$\mathrm{DIC}_1 = -2\langle\log(P(y \mid \theta))\rangle + 2(log(P(y \mid \langle\theta\rangle)) - \langle\log(P(y \mid \theta))\rangle), \tag{3.2}$$

where $\langle\cdots\rangle$ denotes an average over the posterior distribution (note, with a rearrangement, this expression is analogous to AIC in equation (3.1) if we associate $\log(P(y|\langle\theta\rangle))$ with $\log(P(y|\theta_{\max}))$ and

$2[\langle \log(P(y|\theta)\rangle - \log(P(y|\langle\theta\rangle))]$ with effective parameter number p_m). Secondly, Gelman [24] proposed

$$\mathrm{DIC}_2 = -2\langle \log(P(y\,|\,\theta))\rangle + 2\,\mathrm{var}[\log(P(y\,|\,\theta))], \tag{3.3}$$

where now the effective parameter number is given by twice the posterior variance of the observed data likelihood.

3.2. Deviance information criterion for problems with latent variables

Defining a DIC measure in cases when the model contains latent variables is problematic (for example, Celeux *et al.* [8] investigated eight potential definitions). Analogous to equations (3.2) and (3.3), one possibility is simply to use the observed data likelihood as before, but now average over the latent space x:

$$\left.\begin{aligned} &\mathrm{DIC}_3 = -2\langle \log(P(y\,|\,\theta,x))\rangle + 2(\langle \log(P(y\,|\,\langle\theta\rangle_{\theta\,|\,x},x))\rangle_x - \langle \log(P(y\,|\,\theta,x))\rangle)\\ \text{and}\quad &\mathrm{DIC}_4 = -2\langle \log(P(y\,|\,\theta,x))\rangle + 2\,\mathrm{var}[\log(P(y\,|\,\theta,x))]. \end{aligned}\right\} \tag{3.4}$$

Here, $\langle\cdots\rangle$ represents an average over the full posterior, $\langle\cdots\rangle_{\theta\,|\,x}$ is the posterior average over parameters θ, given a particular latent variable state x, and $\langle\cdots\rangle_x$ a posterior average over the latent variable space. A second option considered by Celeux *et al.* [8] is to use the complete posterior probability (i.e. including the contribution from the latent process likelihood as well as the observed data likelihood):

$$\left.\begin{aligned} &\mathrm{DIC}_5 = -2\langle \log(P(y,x\,|\,\theta))\rangle + 2(\langle \log(P(y,x\,|\,\langle\theta\rangle_{\theta\,|\,x}))\rangle_x - \langle \log(P(y,x\,|\,\theta))\rangle)\\ \text{and}\quad &\mathrm{DIC}_6 = -2\langle \log(P(y,x\,|\,\theta))\rangle + 2\,\mathrm{var}[\log(P(y,x\,|\,\theta))]. \end{aligned}\right\} \tag{3.5}$$

There is no consensus on a theoretical justification for a preference between these various options.

3.3. Relationship between deviance information criterion and model evidence

To investigate the relationship between model evidence and DIC, we consider a simple case for which analytical results are derivable (note, understanding this section is not necessary for the rest of the paper, so it may be skipped). Here, we assume a model with no latent variables x and a multi-variate normal (MVN) distribution for the observed data likelihood[5]

$$P(y\,|\,\theta,x) = P(y\,|\,\theta) = P(y\,|\,\langle\theta\rangle)\,\mathrm{e}^{-1/2(\theta-\langle\theta\rangle)^T\Sigma^{-1}(\theta-\langle\theta\rangle)}, \tag{3.6}$$

where θ is the focus of inference, $\langle\theta\rangle$ represents the parameter set corresponding to the maximum observed data likelihood and Σ is a covariance matrix between model parameters, both assumed known. Calculating model evidence requires a proper prior.[6] For simplicity, we choose this to also be MVN and centred on $\overline{\theta}$:

$$\pi(\theta) = \frac{1}{\sqrt{(2\pi)^d\,|\,\Omega\,|}}\mathrm{e}^{-1/2(\theta-\overline{\theta})^T\Omega^{-1}(\theta-\overline{\theta})}. \tag{3.7}$$

The product of two MVNs is also MVN, so equations (3.6) and (3.7) can be multiplied and the parameters integrated out to give (see electronic supplementary material, appendix D for details)

$$P(y) = P(y\,|\,\langle\theta\rangle)\sqrt{\frac{|\,\psi\,|}{|\,\Omega\,|}}\mathrm{e}^{-1/2(\langle\theta\rangle^T\Sigma^{-1}\langle\theta\rangle+\overline{\theta}^T\Omega^{-1}\overline{\theta}-\mu^T\psi^{-1}\mu)}, \tag{3.8}$$

where μ and ψ are the mean and covariance matrix of the posterior. To make a comparison with DIC, it is of use to take minus two times the log of the evidence

$$-2\log(P(y)) = -2\log(P(y\,|\,\langle\theta\rangle)) + \langle\theta\rangle^T\Sigma^{-1}\langle\theta\rangle + \overline{\theta}^T\Omega^{-1}\overline{\theta} - \mu^T\psi^{-1}\mu - \log\left(\frac{|\,\psi\,|}{|\,\Omega\,|}\right). \tag{3.9}$$

[5]Note, in the absence of latent variables, this quantity is often referred to as simply the 'likelihood', but for consistency within this paper, we persist in calling it the observed data likelihood.

[6]A proper prior implies one that can be sampled from. An uninformative (flat) prior over an infinite parameter space is improper.

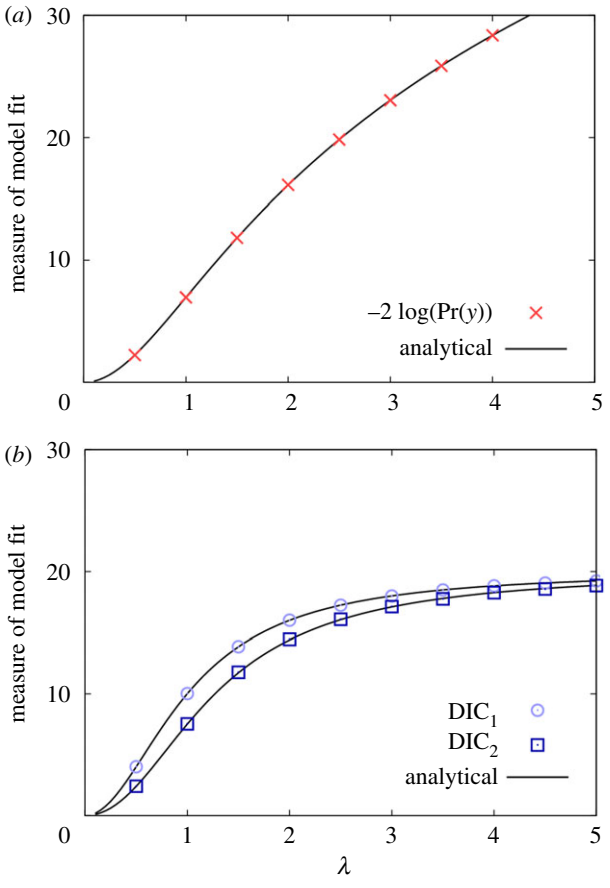

Figure 2. Model with MVN observed data likelihood and prior (and hence posterior). Shows how (*a*) an evidence-based model selection measure and (*b*) two DIC measures (equations (3.2) and (3.3)) vary as a function of the standard deviation in the prior distribution λ.

Furthermore, calculating the posterior mean and variance of the log of the observed data likelihood in equation (3.6) and substituting them into the definitions for DIC in equations (3.2) and (3.3) yields the following analytical results

$$\left.\begin{aligned} \mathrm{DIC}_1 &= -2\log(P(y\,|\,\langle\theta\rangle)) + 2[\mathrm{Tr}(\Sigma^{-1}\psi) + \Delta\theta^T\Sigma^{-1}\Delta\theta] \\ \text{and}\quad \mathrm{DIC}_2 &= -2\log(P(y\,|\,\langle\theta\rangle)) + \mathrm{Tr}(\Sigma^{-1}\psi) + \Delta\theta^T\Sigma^{-1}\Delta\theta + \mathrm{Tr}(\Sigma^{-1}\psi\Sigma^{-1}\psi) + 2\Delta\theta^T\Sigma^{-1}\psi\Sigma^{-1}\Delta\theta, \end{aligned}\right\} \tag{3.10}$$

where Tr denotes the trace of a matrix and $\Delta\theta = \mu - \langle\theta\rangle$ (see electronic supplementary material, appendix D for details).

Note the measures given in equations (3.9) and (3.10) all share a common first term. Thus, comparison naturally focuses on the subsequent terms. We consider analysing a model with $p_m = 10$ parameters under the following scenario: for the observed data likelihood, the diagonal elements of Σ are set to 1 and off-diagonal elements are drawn from a uniform distribution between −0.1 and 0.1, for the prior Ω is diagonal with elements λ^2, both distributions are set to have a common mean $\langle\theta\rangle = \overline{\theta} = \mu$ and we arbitrarily choose $P(y|\langle\theta\rangle) = 1$.

Figure 2 shows how the various model selection measures vary as a function of the prior standard deviation λ. The crosses in figure 2*a* are calculated using SS from equation (2.3) (10^3 burn-in steps and $N = 10^5$ iterations sufficient to generate a high degree of accuracy) and the solid black line shows that they are in excellent agreement with the analytical expression from equation (3.9). The DIC results in figure 2*b* are obtained by running a single posterior chain (10^3 burn-in steps and $N = 10^5$ iterations), and again agreement with the analytical results in equation (3.10) is good.

Figure 2 illustrates notable differences between DIC and evidence-based measures. It is important to emphasize that these differences do not arise from DIC inaccurately approximating the evidence. Rather, the two approaches are simply different measures arising from contrasting philosophies: (i) for model

evidence, a 'good' model is taken to be one that when sampled from gives good agreement with the data (in this case, sampling means drawing parameters from the prior). Therefore, making the prior less and less informative (by increasing λ) results in agreement with the data inevitably going down, and consequently, the measure $-2\log(P(y))$ approaching infinity. (ii) For DIC approaches, a 'good' model is one that contains some set of parameters that make the data likely, and then adds in a penalty term to account for model complexity. In the limit of a flat prior, the penalty terms in equation (3.10) both tend towards $2p_m$, just as for AIC in equation (3.1) (hence explaining why the DIC curves in figure 2 converge on 20, given the model contains $p_m = 10$ parameters).

It should be noted, however, that the differences in figure 2 do not validate one approach over another. This is because when performing model selection, the *absolute* value for the measure is unimportant. It is the *differences* in measure between models that are key. Therefore, validity can only be tested by investigating the ability of these approaches to correctly discriminate within a range of potential models.

Furthermore, it should be stressed that the stated aim of DIC is not to find the 'true' model, but rather to accurately predict future datasets in a world in which the true data generating is, in fact, very high dimensional (and effectively unobtainable) [6,7]. Nevertheless, this paper contends that in situations in which the true model *is* low dimensional and exactly know (as it is for the simulated datasets in §5), it should be expected that DIC does a reasonable job of selecting the true model, since simulation of the observed data from the true model intuitively seems more plausible than from an incorrect one. Therefore, this paper uses model selection, rather than prediction accuracy (or any other measure of model fit, such as posterior predictive assessment [25]), as a means of comparing DIC and model evidence, but concedes that this is not an entirely fair comparison.

4. Relative computational speed

Before assessing model selection using the evidence and DIC, we first briefly investigate the expected computational speeds with which these quantities can be accurately estimated. All the algorithms contain the same fundamental 'update': one in which a complete set of Metropolis–Hastings proposals are applied to allow changes in latent space (typically, this might act on each parameter and latent variable in turn). Updates take approximately the same computational time regardless of method. Therefore, one way to compare computational efficiencies is to establish how many updates U are required to achieve a certain level of sampling variance ε^2 (i.e. noise associated with sampling error) in the corresponding model selection measure. In the case of SS, this measure is taken to be $-2\log(P(y))$ to allow for direct comparison with the DIC results. Assuming large K, the following analytical results for U can be derived (see electronic supplementary material, appendix E for further details):

$$\text{SS}\colon \frac{4}{\varepsilon^2}K\sum_{k=2}^{K} n_k^{\text{cor}}\sigma_k^2(\phi_{k-1}-\phi_k)^2 \quad \text{DIC}_1\colon \frac{16n_1^{\text{cor}}\sigma_1^2}{\varepsilon^2} \quad \text{DIC}_2\colon \frac{8n_1'^{\text{cor}}\sigma_1^4}{\varepsilon^2}, \tag{4.1}$$

where n_k^{cor} represents the characteristic number of updates on chain k over which MCMC becomes uncorrelated (also represented by $n_k'^{\text{cor}}$ but with a slightly different definition) and σ_k is the standard deviation in the log of the observed data likelihood on chain k (note $k = 1$ corresponds to the posterior chain used for the DIC calculations).

The distributions for σ_k and n_k^{cor} are problem-dependent, thus making it difficult to definitively say which approach is the fastest; however, the above expressions do enable a broad comparison. One key point to emphasize is that the computational efficiency for SS in equation (4.1) is *independent* of K for large K (if K is doubled then the temperature separation is approximately halved and these contributions cancel each other out).[7] This result is surprising, and goes some way to explain why despite the fact that SS uses $K = 50$ chains in this study, it is comparable in speed to DIC which uses just a single chain $K = 1$.

The expressions in equation (4.1) also contain other notable features: firstly, in implementing SS, it makes sense to have smaller temperature separations at lower inverse temperatures, as this is the region in which the variance is the greatest (e.g. figure 1*b*), thus helping to motivate the temperature selection scheme in equation (2.4). Secondly, SS benefits from swapping between adjacent chains because this helps to reduce n_k^{cor}. Thirdly, as the number of dimensions in the model increases (and/or the prior becomes less informative), the difference in the observed data likelihood between the prior and posterior chains

[7]So an algorithm which uses $K = 100$ chains is no slower than the one which uses $K = 50$ chains. This remains true for larger and larger K until burning-in such a substantial number of chains becomes problematic.

will inevitably go up. This will have the consequence of increasing σ_K relative to σ_1;[8] hence, for very large problems, SS might be expected to be slower than DIC. Lastly, DIC_2 contains an additional factor of σ_1^2 compared to DIC_1. Since σ_1^2 is a measure of model complexity, this implies that DIC_2 is expected to be slower relative to other methods as model size is increased.[9]

5. Assessment using benchmark models

In this section, we use simulated data to compare model selection accuracy and computational speed using the SS and DIC approaches introduced previously. Three different types of model are considered: (i) a linear regression model (which does not have latent variables and has a nearly MVN posterior distribution), (ii) a mixed model (which incorporates latent variables but again is approximately MVN), and (iii) a set of compartmental epidemic models (which are not MVN).

5.1. Linear regression

Consider a set of measurements y_r (where r runs from 1 to R). An example would be the heights of different individuals in a population. The aim of linear regression is to help explain these measurements in terms of certain known factors, or 'regressors', e.g. gender, age, nationality. The model itself is described by

$$y_r = \sum_{j=1}^{J} X_{rj}\beta_j + \varepsilon_r, \tag{5.1}$$

where X_{rj} is a design matrix and β_j are regressors (where j runs from 1 to J). For example, β_2 could represent the average height difference between males and females (in which case, we might set $X_{r2}=0$ for all females in the population and $X_{r2}=1$ for all males). By convention, $X_{r1}=1$ for all r such that β_1 is the intercept (or average value of y_r when all other regressors are set to zero). The residuals ε_r in equation (5.1) account for any discrepancy between the predictions made by the regressors and the actual observations. They are assumed to be normally distributed with variance η^2.

The linear regression model contains parameters $\theta=\{\beta,\ \eta^2\}$ but no latent variables x. From equation (5.1), the observed data likelihood is given by the product of normal distributions

$$P(y\,|\,\theta,x)=\prod_{r=1}^{R}\frac{1}{\sqrt{2\pi\eta^2}}\mathrm{e}^{-(1/(2\eta^2))(y_r-\sum_j X_{rj}\beta_j)^2} \tag{5.2}$$

and the latent process likelihood is simply $\pi(x|\theta)=1$.

In terms of model selection, one of the key difficulties faced by the scientist is determining which regressors are genuine (i.e. actually affect the observations) and which are not. Specifically, suppose data are available from J' potential regressors, where J' is greater than (or equal to) the true number of regressors J. We denote κ as a J' dimensional vector that specifies a particular model, such that $\kappa_j=1$ if the model contains regressor j and $\kappa_j=0$ otherwise. Thus, in total, $2^{J'}$ possible models exist. If J' is small then it may be practical to calculate model selection measures for each possibility, but as J' increases, the number of potential models can become vast. In this case, two approaches can be taken: (i) in a Bayesian setting model selection MCMC [26] can be used to generate posterior samples for the vector κ [26] and (ii) stepwise methods [27] which minimize the model selection measure by accepting or rejecting successive changes to the model (see electronic supplementary material, appendices G and H for details).[10]

The focus of this paper, however, is not to get into the details of how model selection is actually achieved using these two approaches. Rather, a question of more fundamental importance is addressed: Is the model selection measure actually minimized at (or near to) the true model? (Note, this is an example in which it is implicit that DIC should perform well.) We address this question by considering a simple toy example. Here, $J'=10$ regressors are assumed to exist (elements of the design matrix X_{rj} for $j=2\ldots J'$ are set to 0 or 1 with equal probability). The first five of these regressors are assumed to actually influence the trait and the last five have no influence (i.e. the true model is represented by $\kappa_j=1$

[8]A curious property of the graph in Figure 1*a* is that the gradient of the black line is equal to the variance in the log of the observed data likelihood.

[9]This property also holds true when comparing DIC_4 and DIC_6 with DIC_3 and DIC_5 due to the relative inaccuracy of estimating variances compared to means.

[10]It is worth noting that these stepwise methods are complicated when higher order, i.e. nonlinear, interactions in the explanatory variables are included, and under these circumstances, reversible jump MCMC may become the method of choice.

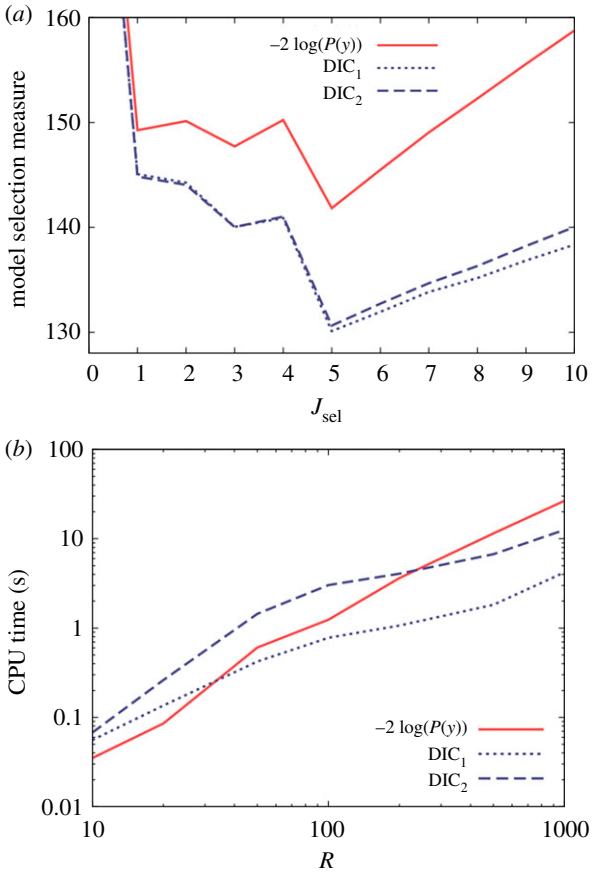

Figure 3. Results from a linear regression model with 10 regressors, the first five of which determine the observed data. (*a*) Model selection measures based on models with a varying number of regressors J_{sel}, where only those regressors J_{sel} and below are included (hence, $J_{sel} = 5$ represents the true model). Here, $R = 100$. (*b*) CPU time necessary to accurately estimate the model selection measures (within an uncertainty of 0.2) as a function of the number of observations R.

for $j = 1 \ldots 5$ and $\kappa_j = 0$ for $j = 6 \ldots 10$). $R = 100$ measurements are simulated from the true model by means of equation (5.1) assuming $J = J'$, $\beta_j = 0.5$ for $j = 1 \ldots 5$, $\beta_j = 0$ for $j = 6 \ldots 10$, and residual variance is taken to be $\eta^2 = 1$.

Next, we consider a subset of potential models which are characterized by the number of regressors J_{sel} they contain and defined such that $\kappa_j = 1$ for $j \leq J_{sel}$ and $\kappa_j = 0$ for $j > J_{sel}$ (note here that $J_{sel} = 5$ represents the true model). The priors for β_j are taken to be uniform between -2 and 2, and for η^2 uniform between 0.1 and 2 (these bounds ensure the prior is proper and sufficiently diffuse to have a negligible effect on the posterior distribution as compared to using an infinite flat prior).[11] Figure 3*a* shows the various model selection measures as a function of J_{sel}. For all the results in this paper, 2×10^3 burn-in updates are used, followed by $N = 10^5$ sampling updates in the case of SS and $N = 10^6$ updates in the case of the DIC (the details of the MCMC implementation are provided in electronic supplementary material, appendix J). The results using model evidence are shown by the solid red line. Those models with $J_{sel} < 5$ are missing certain factors that help to explain the data, so naturally they are not expected to be as good (i.e. there is less evidence supporting them, so $-2\log(P(y))$ is higher). Those models with $J_{sel} > 5$ contain extra model parameters not expected to provide any new information (as they were not used in generating the data). Consequently, the solid curve has a long-term upward trajectory towards the right of figure 3*a*. The minimum lies at $J_{sel} = 5$, indicating that model evidence had successfully identified the true model.

The dotted and dashed blue lines in figure 3*a* provide model selection measures using DIC from equations (3.2) and (3.3). They are both similar and in good qualitative agreement with the

[11]Note, a conjugate prior (instead of this flat prior) is often used to increase the speed at which the model evidence can be calculated. As this paper is focused on the generic performance of DIC and model evidence measures, we do not make use of this mathematical 'trick'.

evidence-based results. One notable feature, however, is that DIC does not penalize against redundant parameters as strongly as the evidence (shown by the increasing separation between the blue and red curves from left to right in this diagram).

Next, we investigate the speed with which the various model selection measures can be estimated. This is achieved by running a large number (100) of independent runs of a given inference algorithm (either DIC or SS), and then calculating the CPU time after which the standard deviation in the model selection measure across runs falls below a certain critical value (which is taken to be 0.2). To remove stochastic noise, the procedure is replicated over four datasets with the CPU times being averaged (further details are given in electronic supplementary material, appendix I).

Figure 3*b* shows this CPU time as a function of the number of observations R.[12] We find that DIC_2 is consistently slower than DIC_1, as expected from the expressions in equation (4.1), and model evidence takes a similar amount of time (despite the fact that the SS method runs using 50 chains). The curves exhibit similar scaling properties, although SS clearly becomes slower relative to DIC for larger system sizes, in line with the third point made in §4.

Illustrative annotated code (written in C++) showing the implementation of the various approaches for the linear regression model (as well as the other models below) is available in the electronic supplementary material.

5.2. Mixed model

In mixed models, the so-called 'random effects' u_q (where q runs from 1 to Q) are added to equation (5.1) through a second design matrix Z_{rq}:

$$y_r = \sum_{j=1}^{J} X_{rj}\beta_j + \sum_{q=1}^{Q} Z_{rq}u_q + \varepsilon_r. \tag{5.3}$$

These random effects are assumed to be drawn from a distribution with mean zero and covariance matrix $\mathbf{G}$.

One application of random effects is to explain correlations in traits between genetically related individuals (in the height example used above, a person's height is, to a certain extent, related to the average height of their parents). Here, the following simplifications can be made [28]: $\mathbf{Z}$ becomes the identity matrix with $Q = R$ (implying one random effect per individual), and $\mathbf{G} = \omega^2\mathbf{A}$ (where $\mathbf{A}$ is an $R \times R$ 'relationship matrix' that captures relatedness between individuals in the population). The relative size of genetic to residual effects (commonly termed environmental effects in this context) is captured through the heritability

$$h^2 = \frac{\omega^2}{\omega^2 + \eta^2}. \tag{5.4}$$

This model contains parameters $\theta = \{\beta, \eta^2, \omega^2\}$ and latent variables $x = \{u\}$. From equation (5.3), the observed data likelihood is given by the product of normal distributions

$$P(y \mid \theta, x) = \prod_{r=1}^{R} \frac{1}{\sqrt{2\pi\eta^2}} \mathrm{e}^{-(1/(2\eta^2))(y_r - \sum_j X_{rj}\beta_j - u_r)^2} \tag{5.5}$$

and the latent process likelihood is MVN

$$\pi(x \mid \theta) = \frac{1}{(2\pi)^{R/2}\omega^R\sqrt{|A|}} \mathrm{e}^{-(1/(2\omega^2))u^T A^{-1}u}. \tag{5.6}$$

To simulate data, we assume a population of 50 individuals with random mating over four generations (making $R = 200$ observations in total), environmental variance $\eta^2 = 0.5$ and heritability $h^2 = 0.5$. As in §5.1, we assume $J = J' = 10$ regressors but that only five of them actually contribute to

[12]Here, four datasets are simulated and the CPU time using the procedure in electronic supplementary material, appendix I, are averaged over.

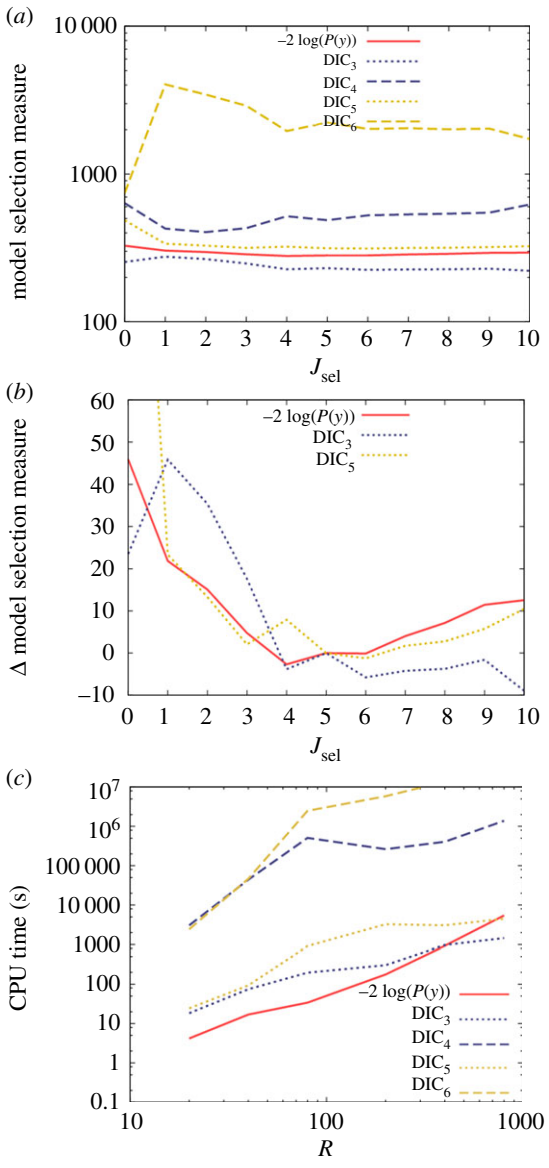

Figure 4. Results from a mixed model with 10 regressors (otherwise known as fixed effects), the first five of which determine the observed data, and one random effect for each individual. (*a*) Model selection measures based on models with a varying number of regressors J_{sel}, where only those regressors J_{sel} and below are included (hence, $J_{sel} = 5$ represents the true model). Here, $R = 200$. (*b*) The model selection measures relative to the true model (for clarity, a table showing these values has been placed into electronic supplementary material, appendix O). (*c*) CPU time necessary to accurately estimate the model selection measures (within an uncertainty of 0.2) as a function of the number of observations R.

the trait (i.e. $\beta_j = 0.5$ for $j = 1 \ldots 5$ and $\beta_j = 0$ for $j = 6 \ldots 10$). The priors for β_j are taken to be uniform between -2 and 2, and for η^2 and ω^2 uniform between 0.1 and 2. Details of how the relationship matrix $\mathbf{A}$ is calculated are given in electronic supplementary material, appendix K, and MCMC proposals are described in electronic supplementary material, appendix L.

Figure 4*a* shows model selection measures for a subset of models characterized by J_{sel} ($\kappa_j = 1$ for $j \leq J_{sel}$ and $\kappa_j = 0$ for $j > J_{sel}$). Because of the latent variables in the model, it now becomes necessary to consider the four definitions for DIC presented in equations (3.4) and (3.5). One immediate conclusion from figure 4*a* is the unreliability of DIC_6 as a model selection measure. It has no minimum near to $J_{sel} = 5$ (the true model) and actually favours a model with no regressors at all. Similarly, DIC_4 is also incorrect because it has a minimum at $J_{sel} = 2$.

Having discounted these two measures, we turn our attention to the other three curves in figure 4*a*. Owing to their large separation, they are difficult to compare on an absolute scale. Model comparison,

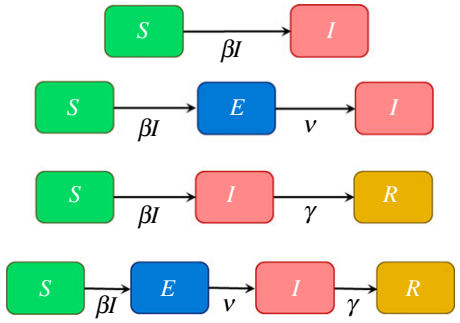

Figure 5. Four different models to describe disease dynamics. Compartments *S*, *E*, *I* and *R* refer to individuals being susceptible, exposed, infectious or recovered, respectively. The arrows indicate transition rates in disease status. (Note, these models are individual-based, so transitions from susceptible to infectious are given on a *per capita* basis βI rather than the usual βSI.)

however, only depends on the relative goodness of fit between models using the same measure. Consequently, figure 4*b* shows each remaining model selection measure relative to its value for the true model. The evidence-based results (solid red curve) exhibit a minimum at $J_{\text{sel}} = 4$, which is actually one less than the true value. However, one must also consider the certainty with which this optimum selection is made. A Bayes' factor of 10 represents strong evidence for one model over another, and this converts to a difference in $-2\log(P(y))$ of 4.6 between models. Therefore, because models with $J_{\text{sel}} = 5$ and 6 regressors are within 4.6 of $J_{\text{sel}} = 4$ in figure 4*b*, there is actually insufficient evidence to statistically differentiate between these possibilities (due to the limited size of the data).

The DIC_5 measure in figure 4*b* follows the same general pattern as the evidence-based approach, but clearly approximations used in deriving DIC introduce a certain level of spurious fluctuation within this curve. While DIC_3 correctly discounts models with too few regressors, it does not sufficiently penalize overfitting additional redundant parameters (which may be an artefact of not including the latent process likelihood in equation (3.4)).

Figure 4*c* shows the computational speeds of the various approaches. Here, the variance-based measures DIC_4 and DIC_6 are considerably slower (due to the difficulty in accurately estimating variances, as illustrated by the additional factor of σ_1^2 for DIC_2 in equation (4.1)). The model evidence measure exhibits similar scaling properties here as when applied to the linear regression model.

5.3. Epidemiological models

We now consider epidemiological models that have a non-MVN latent space. In particular, we consider four commonly used models for disease spread, as illustrated in figure 5. These models are suitable for homogeneously mixing populations but can readily be extended to account for spatial or social structure and other heterogeneities (e.g. [29–31]) and have been used widely to infer the characteristics of disease dynamics and spread. Individuals in the population are classified according to their disease status: *S* represents susceptible, *E* is exposed, *I* is infectious and *R* is recovered. The acronyms used to refer to the models in figure 5 correspond to which of these classifications the model contains (so, in order, these models are SI, SEI, SIR and SEIR, respectively). Model parameters β, ν and γ determine transition rates between these states. Consider a population of $p = 50$ individuals, 49 of whom are initially susceptible and one infected. Figure 6 shows the dynamic variation in the populations within *S*, *E*, *I* and *R* as a function of time when simulating from each of the four models in figure 5 (these simulations were performed using the Doob–Gillespie algorithm [32], as outlined in electronic supplementary material, appendix M).

Next, we consider observations made on the epidemic as it progresses. We assume that periodically all individuals in the population are tested to help identify their disease status. Typically, such diagnostic tests are imperfect, so we define a sensitivity $\text{Se} = 0.8$ (the probability that a truly infected individual, i.e. in the *E* or *I* states, tests positive) and specificity $\text{Sp} = 0.95$ (the probability that a truly uninfected individual tests negative).

Figure 7 gives a diagrammatic representation of the system (here illustrated for the case of an SEIR model). The blue crosses, red pluses and yellow stars represent the times at which individuals become exposed, infectious or recover, respectively. Collectively, these 'events' ξ can be ordered on a single time line indexed by e (which runs from 1 to E).

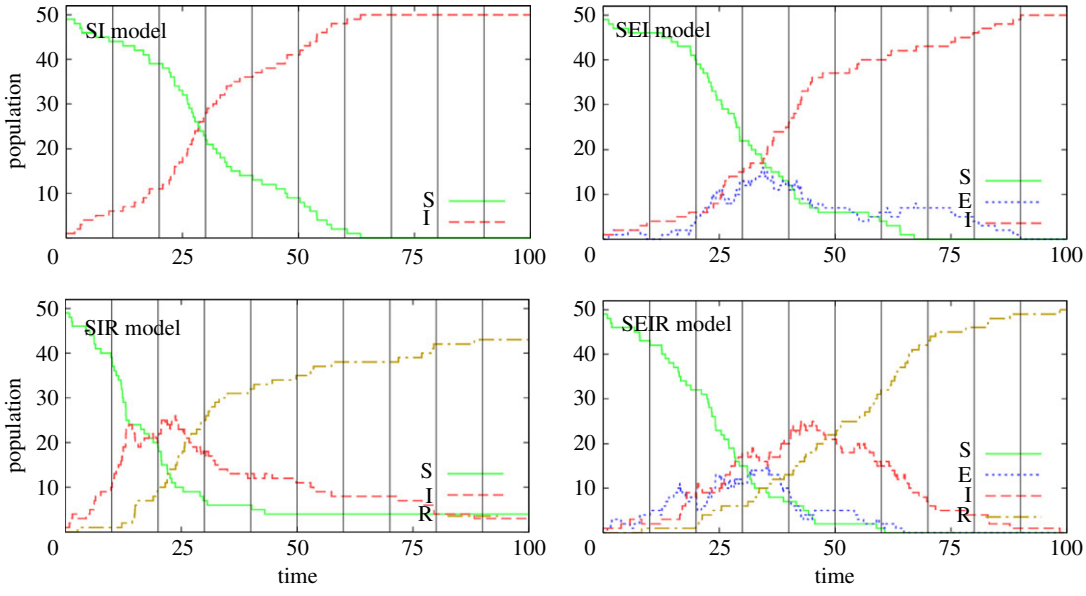

Figure 6. Simulated results from the four models presented in figure 5 (see electronic supplementary material, appendix M for details on how these are generated). Here, we consider a population that initially has 49 susceptible individuals and a single infected. The vertical black lines indicate times at which diagnostic tests are performed. (SI: $\beta = 0.002$, SEI: $\beta = 0.003$, $\nu = 0.1$, SIR: $\beta = 0.004$, $\gamma = 0.05$, SEIR: $\beta = 0.004$, $\nu = 0.1$, $\gamma = 0.05$.)

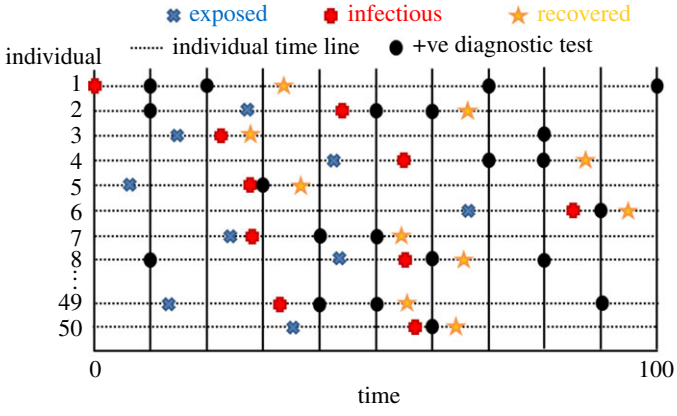

Figure 7. Representation of an individual-based SEIR compartmental model. The horizontal dotted lines represent timelines for individuals within the population, with blue crosses, red pluses and yellow stars denoting different event types (constituting the latent variables x). The vertical black lines are testing times, with positive test results indicated by black circles (note, occasionally positive test results are generated even when an individual is uninfected because the specificity is less than 1).

The events themselves define the latent state $x = \{\xi\}$. The probability of generating this state, given a set of model parameters $\theta = \{\beta, \nu, \gamma\}$, is given by the latent process likelihood

$$\pi(x \mid \theta) = \prod_{e=1}^{E} \rho_{\xi_e} \mathrm{e}^{-W(t_e - t_{e-1})}, \tag{5.7}$$

where t_e and ξ_e are the time and type of event e.[13] In the example of the SEIR model, event rates are given by

$$\rho_{\mathrm{exp}} = \beta I, \rho_{\mathrm{inf}} = \nu, \rho_{\mathrm{rec}} = \gamma, \tag{5.8}$$

(represented by the arrows in figure 5) and $W = S\rho_{\mathrm{exp}} + E\rho_{\mathrm{inf}} + I\rho_{\mathrm{rec}}$ is the total event rate (where S, E and I are the populations of susceptible, exposed and infected individuals immediately prior to time t_e). Note, these expressions are modified dependent on which model is being considered in figure 5.

[13]Here, we set $t_0 = 0$ and, for convenience, assume an end 'event' $t_E = t_{\mathrm{max}}$ with transition probability 1.

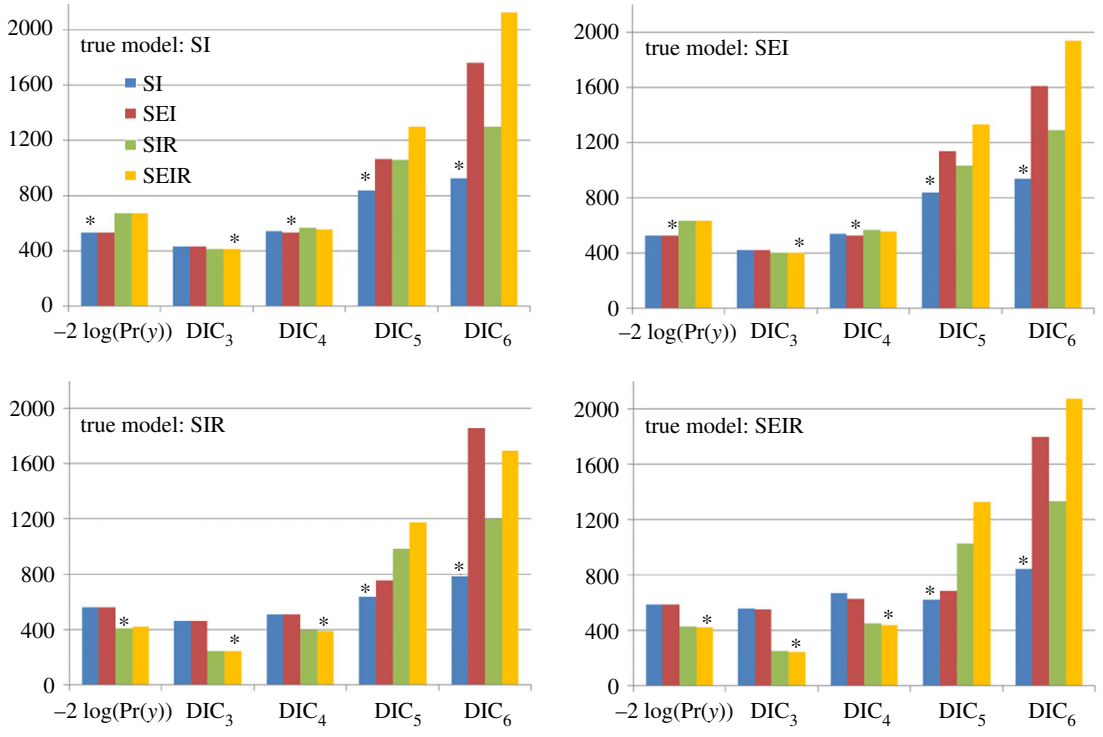

Figure 8. Model selection based on evidence and four DIC measures (equations (3.4) and (3.5)) applied to the simulated results from figure 6. Here, 'true model' refers to the model type used in simulating the data. Out of the potential model which could explain these data, the one with the smallest bar (with star on top) indicates the most likely model based on a given selection measure.

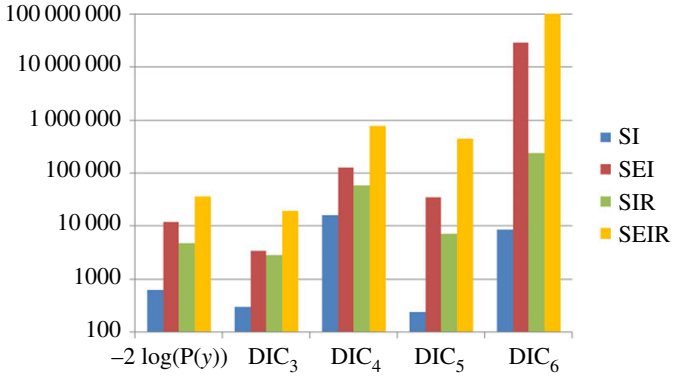

Figure 9. CPU time (in seconds) to accurately calculate the model selection measures (within an uncertainty of 0.2) for $-2\log(P(y))$ (using SS) and the DIC measures defined in equations (3.4) and (3.5).

Disease status testing times are indicated by the vertical lines in figure 7, and the black circles denote those animals which tested positive (otherwise negative). Collectively, these measurements represent the data y. The observed data likelihood is given by

$$P(y \mid \theta, x) = \mathrm{Se}^{N_{+|+}}(1-\mathrm{Se})^{N_{-|+}}\mathrm{Sp}^{N_{-|-}}(1-\mathrm{Sp})^{N_{+|-}}, \tag{5.9}$$

where $N_{r|d}$ is the number of diagnostic tests which give result r from individuals with true disease status d.

Based on the four simulated datasets shown in figure 6, we use model selection measures to attempt to uncover the true model and so reject the other possibilities. Results are shown in figure 8 (details of the MCMC proposals are given in electronic supplementary material, appendix N). Looking first at the evidence-based measure, we find that it selects the correct model in all cases, i.e. $-2\log(P(y))$ is always smallest (indicated by the star) for the true model. Contrast this situation with the DIC results in figure 8,

where in many cases, the 'best' model is incorrect. In fact, all four DIC measures incorrectly identify the true model under at least one scenario.

Figure 9 gives a relative speed comparison between the methods. Again, we find that model evidence is no slower to calculate than the corresponding DIC measures (which, as discussed above, anyway represent poor criteria for identifying the data generating model).

6. Conclusion and discussion

This paper considers two different approaches which can be used to select the best model from a number of candidates: model evidence and DIC.

Model evidence was calculated using SS, which uses samples taken from multiple MCMC chains (run using inverse temperatures that bridge from the posterior to the prior) to construct an estimator. On the other hand, DIC is not uniquely defined and uses samples from a single posterior chain to generate a number of contrasting model selection measures.

The two approaches were compared using a number of benchmark problems (linear regression, mixed models and stochastic compartmental epidemic models). Model evidence was found to consistently predict the correct model. By contrast, the multiple definitions provided by DIC gave different and often contradictory results. While DIC has not been developed with model selection in mind (rather it is concerned with prediction accuracy of future datasets), this lack of consistency between measures (in some cases suggesting completely the wrong model) is of great concern.

Somewhat surprisingly, SS was found to take a similar amount of computational time as DIC, despite having to update a large number of chains. Furthermore, SS is inherently parallelizable, making further gains in speed using GPGPU technology relatively straightforward.

SS provides a robust and practical method for model selection applicable to a wide range of applications, e.g. for hierarchical mixed models used in the estimation of quantitative genetic effects [28] or phylogenetics [33]. It also has important implications for the application of statistical inference to Markov and semi-Markovian compartmental models widely used in epidemiology and ecology, where our results suggest that currently employed model selection approaches based on DIC measures [9–11] are likely to be misleading.

For very large systems estimating evidence using SS will inevitably become computationally demanding. With this in mind, model selection measures other than DIC are currently being developed which use single chain data to provide better model discrimination. One promising possibility is the widely applicable Bayesian information criterion (WBIC) [34]. This calculates the mean in the log of the observation probability at a specially chosen inverse temperature $\phi^* = 1/\log(n)$, where n is the sample size. It has been shown that while WBIC converges on the model evidence in the asymptotic limit [34], i.e. as $n \to \infty$, it can also produce biased results for small sample sizes and when priors are not very informative [35]. It will be interesting to see whether this or something else can be made sufficiently accurate and fast to become the new measure of choice.

Data accessibility. The findings of this paper are theoretical and do not rely on actual data. Exemplar code written in C++ to illustrate the operation of the various computational techniques is included in the electronic supplementary material.

Authors' contributions. C.M.P. wrote the computer code and performed the analysis. C.M.P. and G.M. wrote the manuscript. Both authors gave final approval for publication.

Competing interests. We declare we have no competing interests.

Funding. This research was funded by the Scottish Government through the Strategic Partnership in Animal Science Excellence (SPASE) and the Strategic Research programme of the Scottish Government's Rural and Environment Science and Analytical Services Division (RESAS).

Acknowledgements. We are grateful to Dr Kokouvi Gamado for many helpful discussions.

References

1. Gilks WR, Richardson S, Spiegelhalter DJ. 1998 *Markov chain Monte Carlo in practice*. Boca Raton, FL: Chapman & Hall.
2. Kass RE, Raftery AE. 1995 Bayes factors. *J. Am. Stat. Assoc.* **90**, 773–795. (doi:10.1080/01621459.1995.10476572)
3. Hawkins DM. 2004 The problem of overfitting. *J. Chem. Inf. Comp. Sci.* **44**, 1–12. (doi:10.1021/ci0342472)
4. Green PJ. 1995 Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732. (doi:10.1093/biomet/82.4.711)
5. Hastie DI, Green PJ. 2012 Model choice using reversible jump Markov chain Monte Carlo. *Stat. Neerl.* **66**, 309–338. (doi:10.1111/j.1467-9574.2012.00516.x)

6. Spiegelhalter DJ, Best NG, Carlin BR, van der Linde A. 2002 Bayesian measures of model complexity and fit. *J. R. Stat. Soc. B* **64**, 583–616. (doi:10.1111/1467-9868.00353)
7. Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A. 2014 The deviance information criterion: 12 years on. *J. R. Stat. Soc. B* **76**, 485–493. (doi:10.1111/rssb.12062)
8. Celeux G, Forbes F, Robert CP, Titterington DM. 2006 Deviance information criteria for missing data models. *Bayesian Anal.* **1**, 651–673. (doi:10.1214/06-BA122)
9. Cook AR, Otten W, Marion G, Gibson GJ, Gilligan CA. 2007 Estimation of multiple transmission rates for epidemics in heterogeneous populations. *Proc. Natl Acad. Sci. USA* **104**, 20 392–20 397. (doi:10.1073/pnas.0706461104)
10. Knock ES, O'Neill PD. 2014 Bayesian model choice for epidemic models with two levels of mixing. *Biostatistics* **15**, 46–59. (doi:10.1093/biostatistics/kxt023)
11. Hsu CY, Yen AMF, Chen LS, Chen HH. 2015 Analysis of household data on influenza epidemic with Bayesian hierarchical model. *Math. Biosci.* **261**, 13–26. (doi:10.1016/j.mbs.2014.11.006)
12. Friel N, Wyse J. 2012 Estimating the evidence—a review. *Stat. Neerl.* **66**, 288–308. (doi:10.1111/j.1467-9574.2011.00515.x)
13. Goodman SN. 1999 Toward evidence-based medical statistics. 2: the Bayes factor. *Ann. Intern. Med.* **130**, 1005–1013. (doi:10.7326/0003-4819-130-12-199906150-00019)
14. Jeffreys H. 1961 *Theory of probability*. Oxford, UK: Oxford University Press.
15. Neal RM. 2001 Annealed importance sampling. *Stat. Comput.* **11**, 125–139. (doi:10.1023/A:1008923215028)
16. Xie W, Lewis PO, Fan Y, Kuo L, Chen M-H. 2011 Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Syst. Biol.* **60**, 150–160. (doi:10.1093/sysbio/syq085)
17. Oates CJ, Papamarkou T, Girolami M. 2016 The controlled thermodynamic integral for Bayesian model evidence evaluation. *J. Am. Stat. Assoc.* **111**, 634–645. (doi:10.1080/01621459.2015.1021006)
18. Hug S, Schwarzfischer M, Hasenauer J, Marr C, Theis FJ. 2016 An adaptive scheduling scheme for calculating Bayes factors with thermodynamic integration using Simpson's rule. *Stat. Comput.* **26**, 663–677. (doi:10.1007/s11222-015-9550-0)
19. Friel N, Hurn M, Wyse J. 2014 Improving power posterior estimation of statistical evidence. *Stat. Comput.* **24**, 709–723. (doi:10.1007/s11222-013-9397-1)
20. Friel N, Pettitt AN. 2008 Marginal likelihood estimation via power posteriors. *J. R. Stat. Soc. B* **70**, 589–607. (doi:10.1111/j.1467-9868.2007.00650.x)
21. Casella G, George EI. 1992 Explaining the Gibbs sampler. *Am. Stat.* **46**, 167–174.
22. Roberts GO, Rosenthal JS. 1998 Optimal scaling of discrete approximations to Langevin diffusions. *J. R. Stat. Soc. B* **60**, 255–268. (doi:10.1111/1467-9868.00123)
23. Akaike H. 1998 Information theory and an extension of the maximum likelihood principle. In *Selected papers of Hirotugu Akaike* (eds E Parzen, K Tanabe, G Kitagawa), pp. 199–213. New York, NY: Springer New York.
24. Gelman A. 2004 *Bayesian data analysis*, 2nd edn. Boca Raton, FL: Chapman & Hall/CRC.
25. Gelman A, Meng X-L, Stern H. 1996 Posterior predictive assessment of model fitness via realized discrepancies. *Stat. Sinica* **6**, 733–760.
26. Chipman H, George EI, McCulloch RE. 2001 The practical implementation of Bayesian model selection. In *Model selection* (ed. P Lahiri), pp. 65–116. Beachwood, OH, Institute of Mathematical Statistics.
27. Hocking RR. 1976 A biometrics invited paper. The analysis and selection of variables in linear regression. *Biometrics* **32**, 1–49. (doi:10.2307/2529336)
28. Lynch M, Walsh B. 1998 *Genetics and analysis of quantitative traits*. Sunderland, MA: Sinauer.
29. Lau MSY, Dalziel BD, Funk S, McClelland A, Tiffany A, Riley S, Metcalf CJE, Grenfell BT. 2017 Spatial and temporal dynamics of superspreading events in the 2014–2015 West Africa Ebola epidemic. *Proc. Natl Acad. Sci. USA* **114**, 2337–2342. (doi:10.1073/pnas.1614595114)
30. Keeling MJ, Rohani P. 2008 *Modeling infectious diseases in humans and animals*. Princeton, NJ: Princeton University Press.
31. Brauer F, Castillo-Chávez C. 2012 *Mathematical models in population biology and epidemiology*, 2nd edn. New York, NY: Springer.
32. Gillespie DT. 1977 Exact stochastic simulation of coupled chemical-reactions. *J. Phys. Chem.* **81**, 2340–2361. (doi:10.1021/j100540a008)
33. Posada D, Buckley TR. 2004 Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. *Syst. Biol.* **53**, 793–808. (doi:10.1080/10635150490522304)
34. Watanabe S. 2013 A widely applicable Bayesian information criterion. *J. Mach. Learn. Res.* **14**, 867–897.
35. Friel N, McKeone JP, Oates CJ, Pettitt AN. 2017 Investigation of the widely applicable Bayesian information criterion. *Stat. Comput.* **27**, 833–844. (doi:10.1007/s11222-016-9657-y)